arXiv:1211.3901v1 [cs.CV] 16 Nov 2012

Visual Recognition of Isolated Swedish Sign Language Signs

Saad Akram[1] Jonas Beskow[2] Hedvig Kjellström[1]
[1]CVAP/CAS, KTH, Stockholm, Sweden [2]Speech, Music and Hearing, KTH, Stockholm, Sweden
saadua,beskow,hedvig@kth.se

***Abstract*— We present a method for recognition of isolated Swedish Sign Language signs. The method will be used in a game intended to help children training signing at home, as a complement to training with a teacher. The target group is not primarily deaf children, but children with language disorders. Using sign language as a support in conversation has been shown to greatly stimulate the speech development of such children. The signer is captured with an RGB-D (Kinect) sensor, which has three advantages over a regular RGB camera. Firstly, it allows complex backgrounds to be removed easily. We segment the hands and face based on skin color and depth information. Secondly, it helps with the resolution of hand over face occlusion. Thirdly, signs take place in 3D; some aspects of the signs are defined by hand motion vertically to the image plane. This motion can be estimated if the depth is observable. The 3D motion of the hands relative to the torso are used as a cue together with the hand shape, and HMMs trained with this input are used for classification. To obtain higher robustness towards differences across signers, Fisher Linear Discriminant Analysis is used to find the combinations of features that are most descriptive for each sign, regardless of signer. Experiments show that the system can distinguish signs from a challenging 94 word vocabulary with a precision of up to 94% in the signer dependent case and up to 47% in the signer independent case.**

I. INTRODUCTION

Automatic sign language recognition (SLR) is a challenging research topic that has gained interest rapidly during the past decade. The potential benefits of this technology are obvious: with an ever-increasing information flow in today's society, the sign-language speaking communities are often left to communicate in their second language – the local verbal language. Functional systems for sign language recognition and translation would allow signers to communicate in their first language and be understood by non-signers.

A challenge for SLR systems is also the fact that the same sign can appear differently depending on who is performing it. A signer independent SLR method needs to handle these differences. Fig. 1(a) shows two performances of the sign 'Kan jag få' ('Can I have'). Some signers perform this sign with two hands while others do it with one hand. This can be handled by creating multiple classes for each sign. In addition to that, different signers have different signing styles (how they move their hand and how they make the characteristic hand shapes). Thus, certain aspects of the signing reflect the class of sign being signed, while other aspects reflect the individual style of the signer. An ideal signer independent classifier would ignore the individual style aspects and focus only on class-relevant aspects.

At the same time, two different signs can display very similar visual features, as exemplified in Fig. 1(b-c). Thus, it is important to use as descriptive features as possible. This is a trade-off – low dimensional and signer independent features are robust to intra-class style differences, while high dimensional and rich features are able to pick up on subtle inter-class dependencies.

(a) Two performances of 'Kan jag få' ('Can I have')

(b) 'Rädd' ('Scared') (c) 'Farlig' ('Dangerous')

Fig. 1. Automatic sign language recognition (SLR) from video is a challenging task, due to both high style variation between signers, and very subtle differences between different signs. (a) A sign with high intra-class variation. This is an example of a sign which some signers perform with one hand, and others with two. Moreover, there are style differences between the two signers. (The signer to the left is left handed, something that can be addressed easily by mirroring the video.) (b-c) This is an example of two signs which are difficult to distinguish using hand shape and pose only.

This paper gives four contributions. Firstly, we provide a *method* – to our knowledge the first one in the literature – *for automatic recognition of* a challenging set of *Swedish Sign Language (SSL)* words. Secondly, we introduce a *robust hand segmentation method which employs RGB-D (Kinect) video*. Third, and foremost, we use Fisher Linear Discriminant Analysis (LDA) [1] to *find the most discriminative directions in the feature space across multiple signers*. Finally, we perform sign language recognition from 3D hand trajectories; we *show that the introduction of depth into the hand motion representation increases the recognition rate*.

A. Application: Supportive Signing

The deaf population is not the only group that use signing to communicate. There is a large group of people who use verbal communication but rely on signing as a complement.

Children born with hearing impairment or some form of communication disability such as developmental disorder, language disorder, cerebral palsy or autism, frequently have the need for this type of augmented and reinforced communication. These communication forms are known as TSS ("Signs as Support") in Sweden. They function by "borrowing" individual signs from a signed language (e.g., TSS borrows from SSL). The borrowed signs support and enforce the verbal communication. As such, these communication support schemes do away with the grammatical constructs in sign language and keep only parts of the vocabulary.

One important difference between SSL and TSS is that the latter is poorly formalized and described, and the extent and manner in which it is taught differ widely between different parts of the country. While many deaf children have sign language as their first language and are able to pick it up in a natural way from the environment, children that need signs for other reasons do not have the same rights and opportunities to be introduced to signs and signing. The Swedish TIVOLI project aims at creating a learning environment where children can pick up signs in a game-like setting. An on-screen avatar presents the signs and gives the child certain tasks to accomplish, and in doing so the child gets to practice the signs. The system is thus required to interpret the signs produced by the child and distinguish them from other signs, and indicate whether or not it is the right one and if it was properly carried out. This is a very challenging task due to the large variability that is expected. The sign recognition module should be able to cope with difference in environment, lighting, subject clothing, and subject size. Due to the nature of supportive signing, the system only has to consider the base forms of isolated signs.

The method presented here will serve as the key recognition component of the system. Section III presents the hand segmentation and feature extraction from RGB-D video, Section IV explains how signer independent features are learned and Section V outlines the classification of signs represented by the extracted features. Experiments in Section VI show the method to distinguish signs from a challenging 94 word vocabulary with a precision of 86% on average, 94% for the most skilled signer, when the method was trained and tested on a single signer, and 30% on average, 46% for one signer, when trained on multiple signers and tested on a new signer. Moreover, the introduction of depth into the hand motion representation increases the recognition rate with 10% in the signer dependent case and 25% in the signer independent case.

II. RELATED WORK

We here only focus on non-intrusive video based automatic sign language recognition for dynamic signs; more comprehensive reviews can be found in [2] and [3]. Methods using intrusive data gloves were common at the start of sign language recognition research but in the last decade, vision based methods have become more common and they have started tackling difficult problems like, large vocabularies [4] and sign language recognition in uncontrolled environment [5], [6]. It is common for vision based methods to restrict the background (uniformly colored or static), require the signer to wear full sleeved clothing and even in some cases to wear colored gloves. These restrictions make the task of hand segmentation and tracking significantly easier but at a cost of limiting the system's usability.

Vision based methods rely on multiple image cues to detect and segment hands, these cues include color, motion [7] (frame differencing), edges, background subtraction and region context. Both statistical and adaptive skin color models [8] are common. The adaptive color models can adapt to take different environmental condition (e.g. illumination, signer) in account by changing their model parameters usually using first few or few recent frames in the video sequence.

In the domain of sign language, Kalman filters [7], [8] are the most common method used for tracking hands, used in this paper. [5] used a multiple hypothesis approach and chose the most likely hand trajectory at the end of a sign. Other hand tracking approaches include dynamic programming [9].

Non-manual features, such as head pose and motion [10], [11], facial expression [12], gaze [13], and lip shape [11], convey very useful information in sign language. In the last decade, facial expressions have received the most attention among the non-manual features.

The majority of systems only support signer dependent operation, i.e., every user is required to train the system before being able to use it. Signer independence is usually implemented by some normalization of features, e.g., with respect to the body proportions of the signer [5], or by parametrizing the model by signer identity [14], or training the model with multiple signers [5], an approach that works well if the features are robust [6]. The down-side of robust, crude features are that the precision is lower, which makes the classification task harder.

Most early sign language recognition systems used a form of template matching or neural networks for recognition [15]. However in the last decade or two, Hidden Markov Models (HMM) [16] have become the most common classification method for SLR. Some of the other common methods include Conditional Random Fields (CRF) [17], Dynamic Time Warping (DTW) [18] and nearest neighbor.

III. HAND SEGMENTATION AND FEATURE EXTRACTION

The input to the method is RGB-D (Kinect) video of the signer, along with a 3D estimation of the signer's skeleton from the Microsoft Kinect tracker [19].

Designed for gaming, the Kinect tracker provides the position of hands but this position is not accurate enough to restrict the search space for hands, as can be seen in Fig. 2(a-b). Therefore, a separate hand tracking method (described below) is used to capture hand pose, while the head and torso pose are captured with the built-in tracker.

A. Skin Detection

The signer is required to wear full sleeved, non-skin colored clothing to simplify the hand segmentation. The search

(a) Hands close to body (b) Hands moving fast (c) Hands occluding shoulders

Fig. 2. Examples of error modes in the Kinect tracker. Right (●) and left (●) hand, and shoulder (●) pose estimates are marked.

space for each hand is limited to a small rectangular region in each frame, centered in the Kalman hand pose estimate. Initially the signer is segmented from the background using the Kinect skeleton information.

Human skin color has a restricted range of hue and saturation but there is significant variation in luminance. We thus represent color in terms of a normalized RGB colorspace (($r = \frac{R}{(R+G+B)}$ and $g = \frac{G}{(R+G+B)}$)). Two histogram models (one for skin color and another for non-skin color) are trained using the Compaq dataset [20] These general models are applied to the first frame of each sign in order to get an initial segmentation mask. Skin and non-skin training data, specific to the current signer, are extracted from this mask. These data are used to initialize adaptive, signer specific histogram models, which are in turn updated at each time step t as a linear combination of the models from $t-1$ and t.

The output in each time step t is a binary skin mask S_t.

Another cue to improve the skin detection performance is to use pixel change value. Since among the signer pixels only the hand and arm move frequently, rest of the body remain relatively static. This information can be used to narrow down the pixels which can potentially belong to the hands in any frame. Any pixel with significant change in its value has a higher chance of belonging to the hands. A motion change measure P_t^M is created by taking the image difference of current and previous grey-scale image.

P_t^M is thresholded to a binary motion mask M_t.

The resulting segmentation mask is defined as S_t AND M_t. Then morphological operations (erosion followed by dilation) are applied to remove spurious regions. The resulting potential hand segments are ranked based on their depth (distance from camera), size, and distance from the predicted hand position. The segments with the highest score are assigned to the hands.

B. Hand Tracker

Two Kalman filters are used for each hand; one to keep track of the position of hands (x and y), velocity ($\dot{x}$ and $\dot{y}$), acceleration ($\ddot{x}$ and $\ddot{y}$), the other keeps track of the width and height of the bounding box (w and h) around each hand, and their rate of change ($\dot{w}$ and $\dot{h}$). Some extra padding is added to all four sides of both bounding boxes to accommodate errors in position estimate (when hands change their movement direction or shape abruptly). When hands overlap, this event is detected, and they are treated as one hand.

C. Occlusion Handling

When the predicted bounding box for both hands overlap and there is only one large skin object in the hand search space, hand over hand occlusion's start is marked. During most of the hand over hand occlusion, both hands are either touching each other or are very close to each other. This makes the task of recovering hand shape accurately even with depth information very difficult. Since hand shape usually remains almost same over few consecutive frames in sign language, the hand shape from last frame before occlusion started is used to locate the position of each hand in the joint blob using template matching. During the occlusion, only the position of hands is updated, hand shape features are retained from the last frame before occlusion started.

Another common occlusion in sign language is hand over face occlusion. This occlusion is detected when one of the hand bounding box overlaps the face bounding box. This project solves this type of occlusion using the depth information. A depth model (depth of each pixel) in the face bounding box is created at the start of each sign. This model is updated at each frame. When this occlusion starts, current depth at each pixels is subtracted from depth of each pixel in face model to find out the foreground (hand) pixels. These foreground pixels become the part of the search space for hands. While the occlusion lasts, depth of all pixels excluding those that belong to the hand are updated normally. Fig. 3 shows two frames with hand over face occlusion along with search space for hands after background and face removal and the final segmented hands.

D. Features

The following features are extracted from each frame:

- Image position (x and y) and velocity ($\dot{x}$ and $\dot{y}$) of each hand centroid relative to neck,
- Depth position (z) and velocity ($\dot{z}$) of each hand centroid relative to torso,

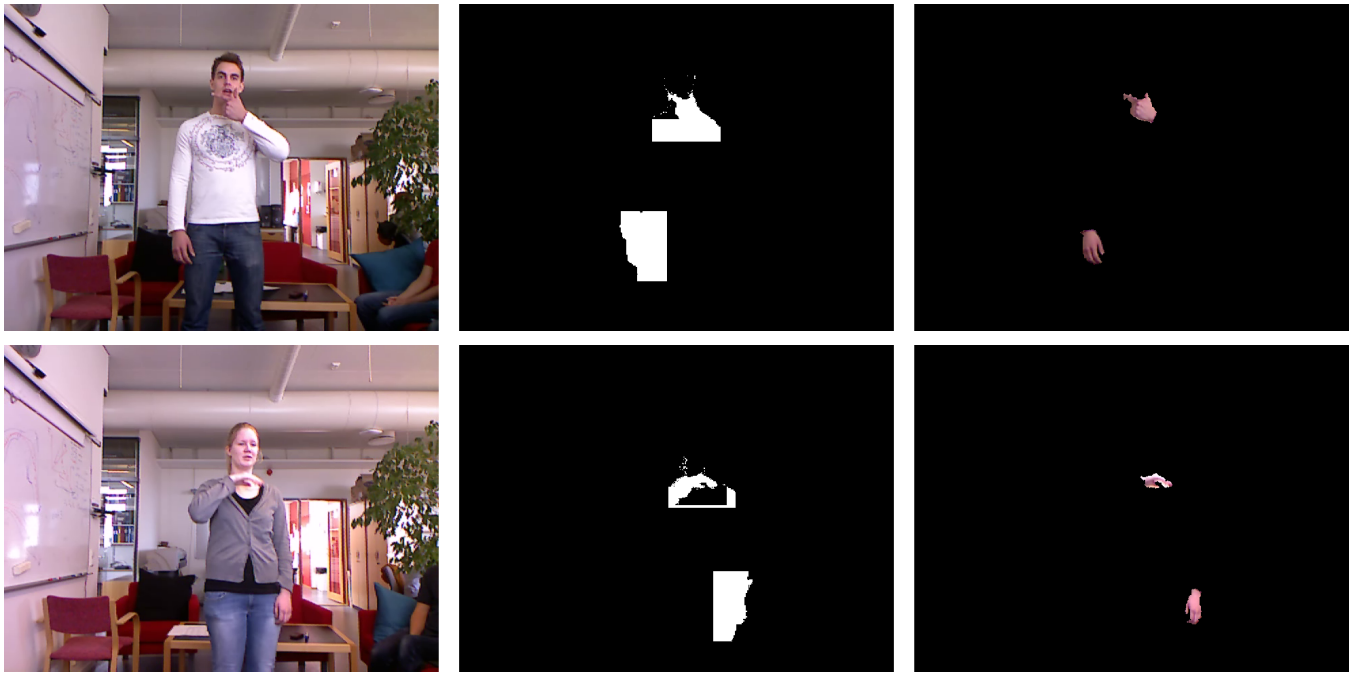


Fig. 3. Two examples of hand segmentation. (left) Frame. (middle) Hands after initial segmentation and face removal. (right) Segmented hands. *For more examples, see the video in the Supplementary Material.*

- Area (a), the number of pixels in the segmented hand which is robust to segmentation errors, and perimeter (p), the number of pixels around the border of the segment which is sensitive to segmentation errors,
- Solidity (s), the proportion of the pixels in the convex hull that are also in the hand segment,
- Major (M) and minor (m) axis length of the ellipse that has the same normalized second central moments as the hand blob, measuring the size of this ellipse,
- Eccentricity ($c = \sqrt{(1 - \frac{m}{M})^2}$) of the ellipse, measuring how much the hand blob deviates from circular,
- Angle ($\cos(\beta)$) between the major ellipse axis and the horizontal image axis,
- Hu invariant moments [21] ($\mathbf{HU}$) which are invariant to image scale, translation and rotation.
- Shape Context [22] ($\mathbf{SC}$). 40 points on the boundary of each hand are chosen, distance and direction to other points is calculated in log-polar space. 5 bins are used for distance (normalized using median distance) and 9 bins for orientation.
- HOG [23] ($\mathbf{HOG}$). Each hand is divided into 2×2 cells and orientation is binned into 9 bin histograms.

Position features are calculated relative to the neck joint of the signer, neck is chosen because it is one of the most stable and accurately tracked joints in the upper torso. Position, perimeter, velocity, and major and minor axes are then normalized using the distance between the signer's shoulders. Distance between shoulders is calculated using the first few frames of each sign, when hands are not likely to be occluding shoulders, something that makes the Kinect tracker compute erroneous shoulder positions, see Fig. 2(c). Area is normalized using the square of the distance between the signer's shoulders. Features for idle hand (the hand which does not move during the whole sign) are set to 0 in a pre-processing step.

We studied the following feature sets (Section VI):

- $\mathbf{posXYZ} = (x, y, z)$
- $\mathbf{posXY} = (x, y)$ from Kalman Tracker
- $\mathbf{posKinect} = (x, y)$ from Microsoft Kinect Tracker
- $\mathbf{velocityXYZ} = (\dot{x}, \dot{y}, \dot{z})$
- $\mathbf{pos} = (x, y, z, \dot{x}, \dot{y}, \dot{z})$
- $\mathbf{S} = (a, p, s, c, M, m, \cos(\beta))$

IV. LEARNING SIGNER INDEPENDENT FEATURES

There is considerable variation between the signing style of different signers as illustrated in Fig. 4. This causes some features (e.g., the horizontal position as in Fig. 4) to vary significantly across signers. HMMs learn the variance of all features for each specific sign, independently of other signs. In the case of Fig. 4, the variance over horizontal position for the signs 'Smaka' ('Taste') and '5' will be large in the HMM states – the horizontal position will then not be taken into regard very much in evaluating a new sign using this model. However, imagine that there is another sign which only deviates from 'Smaka' in terms of horizontal position. Then, the 'Smaka' classifier will give an instance of that sign a high probability of being 'Smaka'.

In contrast, the influence of different features should be decreased or increased based on how discriminative the feature is in separating that sign from other signs. Using Fisher Linear Discriminant Analysis (LDA) [1] – more precisely, its multi-class equivalent – we transform our features to a

new feature space, in which a few dimensions contain most of the discriminatory information. This results in selection of features which are more discriminating across multiple signers and reduces the impact of individual signing style of signers.

Initially feature vectors have to be aligned because the actual signs in different samples s have different start and stop frames. Since we are looking at the difference between different samples at frame level, it is required that all frames within all samples are aligned. We use Dynamic Time Warping (DTW) to achieve this, alignment is done using only position features $\mathbf{posXY}$.

Once all the signs are aligned, they are re-sampled so that they have equal length. The feature vectors from first and last third of the frames are discarded because the motion and hand shape at the start and end of the signs is similar and the middle portion X is most likely to contain the information particular to each sign. At each frame t, the means μ_t^c of each sign c and the total mean μ_t are calculated as

$$\mu_t^c = \frac{1}{N^c} \sum_{n=1}^{N^c} X_{t,n}^c \tag{1}$$

$$\mu_t = \sum_{c=1}^{C} \frac{N^c}{\sum_{c=1}^{C} N^c} \mu_t^c \tag{2}$$

where N^c is the number of samples of class c, and $X_{t,n}^c$ is frame t of sample number n from sign class c.

These mean values from all frames are combined to get the total between-sign scatter matrix S^B and within-sign scatter matrix S^W as

$$S^B = \sum_{t=1}^{T} \sum_{c=1}^{C} (\mu_t^c - \mu_t)(\mu_t^c - \mu_t)^T \tag{3}$$

$$S^W = \sum_{t=1}^{T} \sum_{c=1}^{C} \left(\frac{N^c}{\sum_{c=1}^{C} N^c} \right) \sum_{n=1}^{N^c} (X_{t,n}^c - \mu_t^c)(X_{t,n}^c - \mu_t^c)^T \tag{4}$$

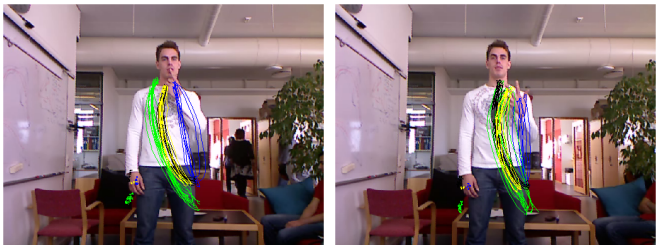
These scatter matrices are used to find the eigenbasis W that maximizes the difference between different signs while minimizing the difference between different samples of same sign. This transformation can be found by solving the characteristic polynomial

$$|S^B - \lambda_i S^W| = 0 \tag{5}$$

to get the eigenvalues λ_i, and then solving for each eigenvector W_i the equation

$$(S^B - \lambda_i S^W) W_i = 0 \tag{6}$$

The M eigenvectors W_i with the highest eigenvalues λ_i are selected, and a signer-independent representation of the features is obtained by projecting them onto this new M-dimensional feature space.



(a) 'Smaka' ('Taste') (b) '5'

Fig. 4. Trajectories for all 4 signers for signs 'Smaka' ('Taste') and '5' : Signer A (●), Signer B (●), Signer C (●), Signer D (●)

V. Sign Classification

HMM [16], which are very common in sign language recognition, were used to represent each sign as a sequence of states in the feature space. In our implementation we used a chain of $N = 7$ active (emitting) states and two (first and last state) non-emitting states, with no skip states. The number of states were determined empirically. The HMM models were trained using the Georgia Tech Gesture Toolkit [24].

VI. Experiments

A. Data collection.

We have collected our own data since our method requires RGB-D video of Swedish Sign Language, which is not available in any of the existing corpora. The vocabulary for this project consists of 94 words, divided into 4 groups ranging in size from 26 to 20. Three of the groups correspond to different games in the tutor system described in the Introduction: *Horror House*, *Fun House*, and *Ice-Cream Stand*, while the fourth, *Numbers*, consists of the numbers 1 to 20. The group *Numbers* consist of many signs which are static and differ from each other only in the number of opened hand fingers.

Some signs are extremely similar in terms of hand motion. An example is shown in Figure 1(b-c): The only difference between the two signs is the facial expression, which indicates that the tutoring application would benefit from the inclusion of non-manual features in the method (see the Conclusions).

Data (two videos and one text file for each signer) were recorded using Microsoft Kinect and Kinect for Windows SDK. Both depth and color videos had a resolution of 640×480 and a frame rate of 30 Hz. The text file contains the location of all joints in the upper body of signer tracked by Kinect. The dataset consists of a total of 23 samples of each sign performed by 4 signers. Three signers (A, C and D) were female and right-handed, while the fourth signer (B) was left-handed and male. For the left handed signer, images and joint positions were flipped horizontally. The number of samples (where each sample contains one instance of each of the 94 signs) recorded were 7 for signer A, 4 for signer B, 5 for signer C and 7 for signer D.

TABLE I

COMPARISON OF DIFFERENT SHAPE FEATURES

	Signer Dependent					Signer Independent				
	mean	A	B	C	D	mean	A	B	C	D
S	77.65%	84.95%	68.88%	78.51%	78.27%	14.14%	14.13%	14.63%	17.02%	10.79%
HOG	83.36%	88.75%	72.34%	85.11%	87.23%	17.32%	21.73%	9.31%	22.13%	16.11%
SC	57.02%	67.48%	45.48%	61.49%	53.65%	15.15%	19.15%	13.56%	15.74%	12.16%
HU	26.95%	25.84%	21.28%	34.26%	26.44%	5.70%	7.60%	6.12%	4.68%	4.41%

All experiments are done with leave one out cross-validation; in the signer dependent case, the classifier is trained on all samples but one for this signer, and tested on the last sample, and in the signer independent case, the classifier is trained on three of the signers and tested on the fourth one.

Setting idle hand features to zero results in significant improvement in the recognition rate when using only shape, or shape and position features and some minor improvement when using position features. This is because it removes the impact of these features which are not needed for recognition but can still significantly lower the probability of a HMM model generating the observed sequence during testing. Min-Max normalization was attempted but it resulted in slightly lower recognition rate compared to not-normalized data and as features are in the same value range, no normalization was used in the experiments.

B. Feature Selection

1) Shape Features: Table I presents the recognition rate for various features used to represent hand shape. The performance of **HOG** features is better than all other shape features in both signer dependent (83.36%) and signer independent (17.32%) tests. One reason is that the HOG descriptor takes the entire hand pattern into account, while the other features are extracted from the silhouette only. Another reason might be that HOGs are not strongly impacted by minor segmentation errors.

Some features in **S** are also quite robust to segmentation errors and slight changes in hand shape, which is why it performs better than **HU** which is sensitive to these sources of noise. The reason for the weak performance of **SC** is probably low image resolution; the hand size in many frames is very small, which means that a small segmentation error can result in a significantly different shape context.

The performance of **S** and **HOG** improves when used in combination with positional features. This is discussed in Section VI-C.

2) Position Features: In this experiment, the aim is to evaluate the amount of information contained in the different pose and velocity features. idle hand features were not set to zero.

Since there is no ground truth of hand positions available, our Kalman tracker was compared to the Kinect tracker by training classifiers with both these sets of features. As can be seen in Table II, the position estimate from our Kalman tracker gives 17% higher recognition accuracy on average, than the same position estimates from the built-in Kinect tracker.

Introducing depth z resulted in an improvement of 10% compared to when only x and y were used in the case of signer dependent recognition and almost same performance in the case of signer independent recognition. When hand shape features are used in addition to position features, the improvement is small; however, this indicates that depth adds robustness to the classifier when color based segmentation fails or can not be attempted (half sleeved clothing, etc).

The signer independent recognition accuracy for signer B and D is significantly lower than signer A and C for all feature combinations. The reason for this is that the trajectories of these two signers deviates in a systematic manner from other signers, as can be seen in Fig. 4. These signers also have higher variance in their trajectories across different samples of same sign. Both these signers were learners of sign language, which can explain the high variability.

C. Signer Dependent Recognition

In this experiment, separate HMM models for signers A, B, C, and D were trained and tested using the **HOG**, **pos**, and **S** feature sets, judged in the previous experiments to perform the best. For each signer, the training was performed on all samples but one, and testing was carried out with the left out sample. All combinations of training and testing samples for each signer were evaluated, and the reported results for each signer A, B, C, and D are the mean of these recognition rates.

Table III lists the recognition rate for all four signers. First of all, it is evident that the feature sets **HOG** and **S** carry

TABLE II

COMPARISON OF POSITION FEATURES

	mean	A	B	C	D
pos (SD)	71.16%	75.08%	70.21%	67.45%	71.88%
posXYZ (SD)	67.86%	73.10%	65.69%	65.32%	67.33%
posXY (SD)	61.83%	69.00%	59.31%	58.09%	60.94%
posKinect (SD)	52.84%	52.13%	50.00%	-	56.38%
pos (SI)	17.47%	22.80%	5.05%	32.77%	9.27%
posXYZ (SI)	13.83%	18.84%	5.05%	28.09%	3.34%
posXY (SI)	13.79%	14.74%	9.31%	27.02%	4.10%

TABLE III

ACCURACY OF SIGNER DEPENDENT RECOGNITION

	mean	A	B	C	D
(**pos**, **S**)	85.06%	91.19%	77.39%	83.19%	88.45%
(**pos**, **HOG**)	85.45%	91.95%	74.73%	87.45%	87.69%
(**pos**, **S**, **HOG**)	87.05%	93.62%	77.13%	87.02%	90.43%

complementary information; the recognition rate improves consistently with 2-3% when they are combined. This makes sense since $\mathbf{S}$ contains information about the area and perimeter of the hand, while $\mathbf{HOG}$ contains complementary information about the shape of the hand, irrespective of absolute scaling.

When studying the recognition using $(\mathbf{pos}, \mathbf{S}, \mathbf{HOG})$, signer A had the highest recognition rate of 93.62% while signer B had the lowest recognition rate of 77.13%.

The reasons for the good recognition of signer A are most probably that she is the most skilled signer of the four, but also that her clothing color is significantly different from her skin color, which gives a very accurate skin segmentation. Fig. 5(a) shows the confusion matrix for signer A. The recognizer confuses the last 20 signs to a higher degree than the first 74. These signs are from the group 'Numbers', which contains signs with very similar hand shapes. Some of the other signs with low recognition rate have another sign with very similar hand shape and trajectory in the vocabulary, e.g. Fig. 1(b-c), which only differ from each other in their facial expression.

One of the reasons why signer B has lower recognition rate than other signers is that there were only 3 training samples for this signer.

D. Signer Independent Recognition after LDA Feature Transformation

In this experiment, 3 signers were used to learn the LDA feature transformation W, described in Section IV. The learned transformation was applied to features from these signers before they were used for training HMM models. When testing, features from the left out signer were transformed using W, and classified using the trained HMM. This was repeated four times, leaving signer A, B, C, and D out for testing in turn. The results are listed in Table IV, and compared with a baseline of performing the training without prior pre-processing using LDA.

TABLE IV

ACCURACY OF SIGNER INDEPENDENT RECOGNITION AFTER LDA FEATURE TRANSFORMATION (NUMBER IN () LISTS # DIMENSIONS USED)

	$(\mathbf{pos}, \mathbf{S})$			
	no LDA(26)	LDA(15)	LDA(20)	LDA(25)
mean	29.04%	34.30%	35.11%	34.94%
A	32.37%	41.64%	42.86%	43.92%
B	16.76%	16.76%	18.09%	16.22%
C	42.55%	46.60%	45.74%	45.74%
D	24.47%	32.22%	33.74%	33.89%
	$(\mathbf{pos}, \mathbf{HOG})$			
	no LDA(84)	LDA(20)	LDA(25)	LDA(30)
mean	23.00%	27.41%	26.66%	27.28%
A	27.05%	37.84%	33.43%	34.50%
B	13.30%	3.72%	4.79%	5.85%
C	29.15%	37.66%	35.74%	34.89%
D	22.49%	30.40%	32.67%	33.89%
	$(\mathbf{pos}, \mathbf{S}, \mathbf{HOG})$			
	no LDA(98)	LDA(20)	LDA(25)	LDA(30)
mean	27.12%	31.40%	31.84%	30.47%
A	29.94%	36.78%	39.97%	36.32%
B	17.82%	4.52%	4.52%	6.12%
C	34.89%	45.11%	44.89%	41.91%
D	25.84%	39.21%	37.99%	37.54%

The mean improvement of LDA is 10-15%. For signer A, C and D, the increase in performance was even higher, 8-37%. The reason for this is probably that the signing styles of these signers deviated from each other in predictable ways – they have all developed an individual, consistent style, just as one develops an individual style for hand writing. However, LDA did not improve the recognition of signer B. The reason is, in analogy to above, that signer B has not yet developed a consistent signing style since he is a learner; his signs vary in a stochastic manner, and the patterns of style variation learned from the other signers do not apply on his signing.

Looking at the confusion matrices in Fig. 5(b) and (c), it is evident that the recognition of some signs has greatly improved by the LDA preprocessing – the signs that show a great style variability between signers – while others are unaffected – the signs that are performed in the same manner independent of signer, or with more random variations.

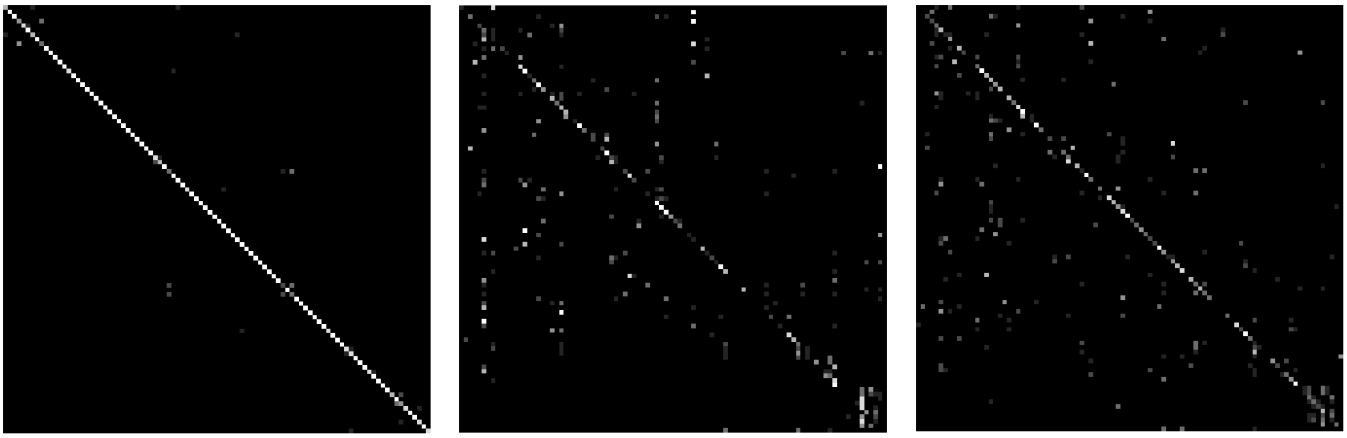
VII. CONCLUSIONS

We present a method for recognizing Swedish Sign Language (SSL) from video. The method will be used in a computer game intended for sign language training for children with communicative disabilities. The signer is captured with an RGB-D (Kinect) system, which gives the possibility to recognize signs in terms of 3D hand motion. The primary contribution of this paper is a method to learn a projection of the hand features using Linear Discriminant Analysis (LDA), which makes the recognition robust to variation between different signers. Additional contributions of the paper are II) the SSL recognition method, III) a robust, hand segmentation method based on color, depth and motion, and IV) the inclusion of hand depth among the classification features.

Since there is no commonly used dataset for sign language recognition, it is very difficult to compare different methods quantitatively. Differences in vocabulary across papers is also very varied, which further complicates the comparison. [5] achieved up to 99% recognition accuracy for signer dependent and 44% in signer independent experiments in controlled environments. Our system has a recognition accuracy of 87% on average in signer dependent experiments and 35% on average in signer independent experiments in a natural environment with a very demanding set of signs.

The present method can primarily be extended in three directions. Firstly, as illustrated in Fig. 1(b-c), certain signs can not be distinguished by manual features alone. We will investigate the inclusion of non-manual features in the recognition.

Secondly, LDA is a linear method for finding the most discriminative directions in feature space. We will investigate non-linear alternatives, such as Gaussian Process Latent Variable Models [25]. This can be expected to improve performance, as the Gaussian assumption on data distribution posed by LDA is probably a simplification of the real data distribution.



(a) Signer dependent (93.62%) (b) Signer independent, no LDA (36.78%) (c) Signer independent (39.97%), LDA(25)

Fig. 5. Confusion matrices for signer A, using features $(\mathbf{pos}, \mathbf{S}, \mathbf{HOG})$.

Moreover, system will be tested using the data from children with communication disabilities.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. H. Stork, Eds., *Pattern Classification*. Wiley Interscience, 2000.

[2] S. Ong and S. Ranganath, "Automatic sign language analysis: a survey and the future beyond lexical meaning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 873–891, 2005.

[3] H. Cooper, B. Holt, and R. Bowden, "Sign language recognition," in *Visual Analysis of Humans*, T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds. Springer, 2011, pp. 539–562.

[4] H. Wang, R. Stefan, S. Moradi, V. Athitsos, C. Neidle, and F. Kamangar, "A system for large vocabulary sign search."

[5] J. Zieren and K.-F.Kraiss, "Robust person-independent visual sign language recognition," in *IbPRIA*, 2005.

[6] H. Cooper, E.-J. Ong, and R. Bowden, "Give me a sign : A person independent interactive sign dictionary," University of Surrey, Guildford, UK, Tech. Rep. VSSP-TR-1/2011, 2011.

[7] K. Imagawa, S. Lu, and S. Igi, "Color-based hands tracking system for sign language recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

[8] J. Han, G. Awad, and A. Sutherland, "Automatic skin segmentation and tracking in sign language recognition," *Computer Vision, IET*, vol. 3, no. 1, pp. 24–35, 2009.

[9] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney, "Tracking using dynamic programming for appearance-based sign language recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2006.

[10] U. M. Erdem and S. Sclaroff, "Automatic detection of relevant head gestures in American Sign Language communication," in *IAPR International Conference on Pattern Recognition*, 2002.

[11] U. von Agris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," in *IEEE International Conference on Automatic Face Gesture Recognition*, 2008.

[12] T. D. Nguyen and S. Ranganath, "Tracking facial features under occlusions and recognizing facial expressions in sign language," in *IEEE International Conference on Automatic Face Gesture Recognition*, 2008.

[13] L. Muir, I. Richardson, and S. Leaper, "Gaze tracking and its application to video coding for sign language," in *Picture Coding Symposium*, 2003.

[14] U. von Agris, C. Blömer, and K.-F. Kraiss, "Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, MLLR, and MAP," in *IAPR International Conference on Pattern Recognition*, 2006.

[15] T. Starner, J. Weaver, and A. Pentland, "Real-time American Sign Language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[17] H.-D. Yang and S.-W. Lee, "Robust sign language recognition with hierarchical conditional random fields," in *IAPR International Conference on Pattern Recognition*, 2010.

[18] P. Doliotis, A. Stefan, C. Mcmurrough, D. Eckhard, and V. Athitsos, "Comparing gesture recognition accuracy using color and depth information," in *Conference on Pervasive Technologies Related to Assistive Environments*, 2011.

[19] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

[20] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[21] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[22] S. Belongie and J. Malik, "Matching with shape contexts," in *IEEE Workshop on Content-based Access of Image and Video Libraries*, 2000.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[24] T. Westeyn, H. Brashear, A. Atrash, and T. Starner, "Georgia tech gesture toolkit: supporting experiments in gesture recognition," in *IEEE International Conference on Multimodal interfaces*, 2003.

[25] J. D. Lawrence, "Gaussian process latent variable models for visualization of high dimensional data," in *Advances in Neural Information Processing Systems 16*, 2004.